Open Economies Review
https://doi.org/10.1007/s11079-023-09723-8

RESEARCH ARTICLE

A Match Made in Maastricht: Estimating The Treatment Effect of the Euro On Trade

Joseph Kopecky[1]

Accepted: 12 June 2023

Abstract
Why do estimates of the European Monetary Union (EMU) effect on trade vary so greatly? Rose (2017) shows that the largest factor determining the size of EMU trade estimates is the choice of sample, with studies using only European or rich countries finding smaller impacts than those using more complete trade datasets. I push this question one step further, asking instead: what is the appropriate comparison group with which to study the euro's trade impact? Using a first stage estimation of selection into the EMU and a robust propensity score weighting estimator, I extend the work of Millimet and Tchernis (2009) to a larger dataset of countries and years, showing that gravity estimates of the euro effect on trade are smaller when sample truncation and weighting brings the differences in observable characteristics between EMU and non-EMU pairs close to zero. Utilizing a Poisson pseudo-maximum likelihood approach, I find that estimates using this more robust estimator reflect the same pattern, but with significantly less initial upward bias. My work suggests that policy analysis in trade should be more careful to consider the comparability of "treated" and "control" observations, and more readily utilize propensity score methods as a data driven approach to rebalancing samples when differences across these groups are large.

JEL F13 · F14 · F15 · F33

1 Introduction

The euro represents a historic policy experiment. There is little precedent for such a large number of wealthy countries to multilaterally surrender control of their currencies. In their assessment of the challenges facing the eurozone in the depths of the

Special thanks to the editor, George S. Tavlas, and to three anonymous referees for their insightful comments and suggestions.

✉ Joseph Kopecky
jkopecky@tcd.ie

[1] Trinity College, Dublin, Ireland

Published online: 20 July 2023

European sovereign debt crisis, O'Rourke and Taylor (2013) show that most prior currency unions make for a poor comparison to the European Monetary Union (EMU). It should not then be surprising that estimates of the impact of the EMU on trade differ so dramatically from those studying these earlier unions. Assignment into the EMU, to borrow language from experimental studies: the policy *treatment*, is not remotely random. An important consideration is how accurately the *control* group (non-EMU country-pairs) compare on observable characteristics to those that receive policy treatment. In recent work, Rose (2017) surveys the literature of EMU trade effects, finding that the driving factor of whether a study finds large or small estimates comes down to sample choice. Rose (2017) advocates that more data should be better, which would suggest that the trade effect is big. I show that sample selection is indeed crucial to determining the size of EMU coefficients, but that studies which reduce the sample[1] may at times offer an improvement in the comparability of these groups. I propose a data driven, propensity score, method of rebalancing the EMU and non-EMU samples to better match on observed characteristics. This not only provides an empirically driven way to select the sample, but also utilizes a robust weighting estimator that leverages gains from matching estimators, while still using well studied, and theoretically grounded, gravity equations of trade to estimate treatment effects.

The effect of currency unions on trade has been hotly debated. Empirical work stems from Rose (2000), whose gravity equation estimates suggest that a common currency more than doubles bilateral trade between countries. This was, by the author's own admission, unreasonably large but surprisingly robust (Rose 2002). A cottage industry sprung up to minimize the currency union effect with Nitsch (2002) suggesting a weaker effect, and Glick and Rose (2002) showing that a smaller, though still extremely large estimate survives limiting estimation to the time-series *within-pair* variation. More recently attention has focused on estimating these effects for the euro-area, with Glick and Rose (2016) providing an update on their earlier currency union estimates to include the reasonably large time-series of euro data. They find an estimate of a roughly 54% increase in trade from EMU membership. This is much larger than the effect found by many others in the literature. A meta analysis carried out in Polák (2019) finds estimates of an EMU trade effect in the 2%-6% range, with more recent evidence pointing to little, if any impact at all. Rose (2017) suggests that much of the difference among EMU trade estimates comes from sample selection, with effects that are increasing as new years are added to the sample, and smaller when limiting the sample to subsets of rich countries (roughly 12% gains) rather than the full bilateral export data as in Glick and Rose (2016). Notably my preferred estimates are closely in line with Kelejian et al. (2012), who estimate the effect of euro on trade on a relatively homogenous sample and accounting for the spatial and persistent nature of trade.

Larch et al. (2019) reproduce the results of Glick and Rose (2016) using the Poisson pseudo-maximum likelihood (PPML) estimators suggested in Silva and Tenreyro (2006) and Silva and Tenreyro (2011), who show that log specifications of the gravity

[1] For example, by including only high income countries

produce biased estimates in the presence of heteroscedasticity. Larch et al. (2019) show that this heteroscedasticity is particularly important in the context of EMU trade estimates. Notable for my work, they find that inclusion of small countries in the dataset produces sizeable impacts on estimates of the EMU effect, amplifying the difference between PPML and OLS results. They find, similar to Rose (2017), that estimates of the currency union and EMU effects vary substantially when altering the sample (OECD, Upper income, etc). My results will echo this conclusion, while contributing to their findings by providing a data driven method of sample selection, chosen to reduce the differences in observable characteristics between treated and control subpopulations. In addition to this, I find that the PPML estimator provides much more stable estimates when the underlying sample changes, further motivating its use relative to log-gravity estimators when problems of sample selection are large.

I am not the first to apply propensity score methods to the currency union effect on trade, nor even to the case of the EMU. Persson (2001) shows that using matching estimators to estimate the average treatment effect of currency unions significantly reduces their estimated impact. Similar matching estimators are used by Chintrakarn (2008) in the context of the euro area, again reducing estimators. Their work relies on estimation of selection into *treatment* groups, using logit and probit models of probability of being in a currency union (or the EMU) and then comparing the conditional mean of treated observations with that of control groups with similar likelihood of being treated. While these estimates work to solve the selection problem that is endemic in macro policy estimates of trade, they also fail to leverage the usefulness of the well studied gravity equations used in work such as Glick and Rose (2016). I instead use a *doubly robust* estimator, combining the propensity score weighting used in work such as Persson (2001) and Chintrakarn (2008) with calculations of the conditional mean using a gravity equation approaches that are standard to the trade literature. My methodolog is in a class of doubly robust estimators, which model both selection into treatment as well as the policy effect on outcomes. Such methods are described in detail in: Imbens (2004), Lunceford and Davidian (2004), and Wooldridge (2007); with applications in the context of macroeconomic policy (fiscal shocks) in Jordà and Taylor (2016).

Millimet and Tchernis (2009) applies my preferred estimator, inverse propensity score weighting with regression adjustment (IPWRA), to the EMU. Their work is primarily interested in providing guidance on specification of the first-stage model, showing that over-fitting such models can be beneficial. Their EMU application studies the period of 1999-2002 and focuses on 22 developed countries. They find an impact in line with similar panel OLS estimates, such as Micco et al. (2003), with a roughly 12% increase in trade due to EMU membership. My estimation updates their results to a much larger trade dataset, while contextualizing the use of the IPWRA as a method of improving on the problem of selection into EMU membership. This latter point is important as the IPWRA estimation procedure is not only a useful estimator, but through the first stage estimation provides a clean empirical way to demonstrate the appropriateness of the sample used in estimation, something ignored in much of the trade literature. Kopecky (2023) uses similar propensity

score methods to show that disaggregated estimates of currency union trade effects.[2] At present sample choice across this literature appears to be completely arbitrary, with Micco et al. (2003), and many others, restricting their estimation sample to developed countries, while other authors make a similar appeal to that of Rose (2017)–more data is better.

The desire to use the largest amount of data is understandable. However, problems of selection bias in estimates of currency unions are well known. It was precisely this issue that motivated Persson (2001) to use matching estimators. While Rose (2001) makes compelling arguments for why the pure matching approach has flaws, the estimator I suggest utilizes the strengths of both these propensity scores and traditional gravity approaches. Knowing the potential issues of selection, increasing the sample size to include observations that have little in common with EMU members appears to bias estimates upwards. This works in the same way that including large pools of healthy, low risk, individuals as a comparison group may bias observational estimates on the efficacy of a medical treatment downward. Their better health makes them less likely to receive the treatment, but lack of treatment has no causal bearing on their health outcomes. The experimental ideal would use randomization to compare those receiving the medical treatment to individuals who need such treatment, but do not receive it. Of course this is not possible with observational data. It is precisely such contexts for which the IPWRA style estimators (and earlier matching estimators) were developed.[3] While I do not claim that this estimator removes all potential issues of selection in the context of the EMU, improving the comparability of treatment and control observations should only work to reduce such problems as much as possible given the observable characteristics available in standard bilateral trade data.

Beyond having an academic methodological interest, the answer to this question is critical. The global financial crisis highlighted many of the challenges and costs associated with eurozone memberhsip. Aizenman (2018) highlights some of these challenges for the euro and other developing market currency unions. The recent inflationary episode will likely prove another hurdle. Trade is only one benefit of the euro to member states, but it will be important for policy going forward to understand which of the diverse set of estimates of the trade impact of the euro and other currency unions accurately reflect the scale of that benefit. My work suggests trade impacts that are small, but still positive, while remaining statistically and economically meaningful.

2 Data and Methodology

I use the CEPII gravity dataset from Conte et al. (2021), constructing a measure of EMU currency union membership consistent with existing measures from Glick and Rose (2016), but excluding some small French territories that are included in their

[2] This primarily focuses on developing countries, but includes the EMU for comparison, showing that aggregate measures of a currency union effect mask significant heterogeneity.

[3] See for an example using IPWRA to study the effectiveness of right heart catheterization, Hirano and Imbens (2001)

analysis.[4] In all results below, I drop non-EMU currency union pairs to ensure that my baseline comparison group is not polluted with countries currently in a different currency union pair that may similarly affect their trade. Comparing to Rose (2017), my full dataset includes more years (covering 1948-2019), but fewer country-pairs. When restricting to log-exports, as in their analysis, I have 29,394 country-pairs compared to their 34,104. The reduction in country-pair observations is in part because I drop non-EMU currency union observations, but also because the source data from Conte et al. (2021) has slightly different coverage than the IMF-DOTS used in their analysis. As a result, initial estimates using the Glick and Rose (2016) model are close-to, but slightly different from their results.

My starting point for estimation is the gravity specification suggested in Head and Mayer (2014). This is the "theory-consistent" method of specifying this empirical relationship, accounting for the multilateral resistance terms that are important in structural gravity equations. To accomplish this I include a full set of exporter-year and importer-year dummy variables. In addition to these, I include a time-invariant set of country-pair fixed effects, which have been shown by many to be important in the context of currency unions, and combined make up the preferred specification in Glick and Rose (2016). This is given by:

$$ln(X_{ijt}) = \gamma EMU_{ijt} + \beta RTA_{ijt} + \lambda_{it} + \psi_{jt} + \phi_{ij} + \epsilon_{ijt} \tag{1}$$

where X_{ijt} are exports from country i to country j at time t, EMU_{ijt} is a dummy variable representing a EMU membership for a country-pair in year t, RTA_{ijt} is control for regional trade agreements, λ_{it} exporter-time fixed effects, ψ_{jt} importer-time fixed effects, and ϕ_{ij} time-invariant country-pair fixed effects. While I can include a large set of standard gravity controls, nearly all are inestimable while using exporter/importer-year and pair fixed effects, so I omit them from discussion here. Including the few with enough variation to remain in this specification have no bearing on estimates of $\hat{\gamma}$.

Silva and Tenreyro (2006) provide a now well-known critique of Eq. (1), showing that under heteroskedasticity, the log-linearized model of trade leads to biased estimates. They show in both Silva and Tenreyro (2006) and Silva and Tenreyro (2011) that estimations of Eq. (1) on trade data suffer substantially from this bias, suggesting instead to use a Poisson pseudo-maximum likelihood (PPML) estimator without taking logs. This methodology has the advantage of not only dealing with the bias introduced by the log-linear approximation, but also allows for inclusion of zero trade flows. While I will continue to use the log gravity specification to link my work to the results of Rose (2017). I also provide estimates of the equivalent PPML estimation, of:

$$X_{ijt} = e^{\gamma EMU_{ijt} + \beta RTA_{ijt} + \lambda_{it} + \psi_{jt} + \phi_{ij}} + \epsilon_{ijt} \tag{2}$$

where controls and fixed effects are defined identically to those above.

[4] As in their work the start of euro membership is the date of adoption, which is 1999 for the eleven initial member states, adding Greece in 2001, Slovenia in 2007, Cyprus and Malta in 2008, Slovakia in 2009, Estonia in 2011, Latvia in 2014, and Lithuania in 2015.

2.1 Inverse Propensity Score Weighting: A Doubly Robust Estimator

Propensity scores feature heavily in the rigorous debate around the size of the currency union effect on trade. In his original rebuttal of the eye-popping estimates of Rose (2000), Persson (2001) showed that matching estimators substantially reduce the impact of the currency union effect on trade. His methodology uses two estimators: *nearest-neighbor* matching, and *stratification*. Both are two-step estimators that rely on a first stage estimate of the probability of selection into a currency union, and then generate estimates for the average treatment effect using a difference in means among these groups. The nearest neighbor method pairs each treated observation with its most comparable control group, while stratification bins treated and control observations according to their probability of treatment and takes a weighted difference in means within each bin. The goal of these methods is to deal with concerns of selection, well understood in the context of currency unions and the euro area.[5] Similar methods are used in a more recent estimate of the euro area in Chintrakarn (2008), who again suggests a downward revision of eurozone estimates after applying matching estimators.

In an initial response to Persson (2001), Rose (2001) correctly critiques the probit model used for selection into currency unions as a poor fit for the data. Histograms of treated (currency union) probabilities reveal that most of the weight of predicted likelihood of being in a currency union falls near bottom of the distribution. Matching estimators such as those used in Persson (2001) rely on the identification of the model of selection into treatment for the second stage difference-in-means estimators. Identification hinges on converting observational data into something closer to a randomized control trial under the assumption that the rebalancing of treatment and controls will ameliorate selection issues. However, the lack of fit and out-of-sample predictions of currency unions should give caution to those relying on the soundness of this identification, especially when comparing their results to the well studied, and theoretically motivated, gravity equation. The case that Rose (2001) makes against their use is rather convincing, particularly when comparing to the modern specifications of gravity that control for time-varying country-specific multilateral resistance terms, as Glick and Rose (2016) does.

However, such critiques of the chosen matching estimators do not address the concerns of selection that motivated the work of Persson (2001) and Kenen (2002). Perhaps it is possible to have our cake and eat it too? A different class of estimators use propensity score matching to estimate *doubly robust* estimators, that replace the difference-in-means second stage with other second stage estimates of conditional means. I make use of inverse propensity score weighting with regression adjustment (IPWRA). This process involves specifying a first-stage treatment estimation of selection into treatment (the EMU), but instead of a difference in means approach,

[5] See for example, Baldwin and Taglioni (2007) who discuss this issue in the context of estimating the euro effect on trade and Baldwin (2006) who compares this problem to an attempt to study the impacts of dieting on weight gain without understanding the underlying model whereby individuals decide to go on a diet.

I then estimate a second stage model using regression specifications with propensity scores serving as weights. This allows for estimates to be rebalanced using first stage estimates similar to that of Persson (2001), while still using the well studied gravity equations to estimate conditional mean across treated and control sets. The *doubly-robust* nature of the IPWRA estimator, is shown and discussed in great detail in work such as Imbens (2004), Lunceford and Davidian (2004), and Wooldridge (2007), and suggests that the estimator will yield consistent estimates of the average treatment effect if *either* model is correctly specified. I follow similar notation to Jordà and Taylor (2016), who develop this estimator in a macroeconomic context.[6] The IPWRA estimator is given by:

$$\widehat{ATE}_{IPWRA} = \frac{1}{n_1^*}\sum\left[\frac{EMU_{ijt}\left(m_1\left(Z_{ijt},\hat{\beta}\right)\right)}{\hat{p}_{ijt}}\right] - \frac{1}{n_0^*}\sum\left[\frac{(1-EMU_{ijt})\left(m_0\left(Z_{ijt},\hat{\beta}\right)\right)}{1-\hat{p}_{ijt}}\right] \tag{3}$$

where EMU_{ijt} is a dummy representing whether a country-pair is in the eurozone at time t, and $\hat{p}_{ijt}$ is the first-stage estimate of probability of treatment. A general term for any second stage estimate of conditional means for treated (1) and control (0) groups is given by: $m_{0/1}\left(Z_{ijt},\hat{\beta}\right)$.[7] I estimate this conditional mean using Eqs. (1) or (2), where $\hat{\beta}$ now represents the vector of all estimated regression coefficients. Estimates of $\hat{\beta}$ can in principle be estimated separately for treated and control populations, as described Sloczyński and Wooldridge (2018), however this is not possible while using the fully specified theory consistent gravity equation described by Head and Mayer (2014), as specifying the conditional mean across treatment and controls requires a high dimension of fixed effects, larger than the number of EMU observations in the sample. As such, I am limited in this application to the assumption that coefficients for estimation of these conditional means are identical, which is of course also used in traditional gravity specifications. In keeping with suggestions of made in Hirano and Imbens (2001) and Imbens (2004) these weighted averages are normalized, with $n_1^* = \sum \frac{EMU}{\hat{p}}$ and $n_0^* = \sum \frac{1-EMU}{1-\hat{p}}$, to ensure that the probability weights are normalized to sum to one.

In practice, estimation of Eq. (3) is quite straightforward. Upon obtaining first stage estimates of $\hat{p}$ inverse propensity score weights are assigned as $\frac{1}{\hat{p}}$, to treated (EMU) observations, and $\frac{1}{1-\hat{p}}$ to controls (non-EMU). The normalization described above simply divides these weights by the sum of all weights to ensure that they sum to one. The second stage requires any conditional mean estimate $m_{1/0}(Z_{ijt},\hat{\beta})$, which for my purposes is the coefficient estimate of the EMU in the gravity equation. Note that I assume that $\hat{\beta}$ is the same for treated and control groups. This assumption allows me to run a single gravity equation with the EMU dummy using the inverse

[6] They use local projections to study fiscal austerity shocks rather than the static gravity equation I employ here.

[7] To condense notation I refer to the entire control set as simply Z_{ijt} with $\hat{\beta}$ referring to all estimated parameters, this includes the rich set of fixed effects in Eq. (1).

propensity scores as weights.[8] It is possible to allow for separate estimation of the conditional means, (ie $m_1(Z_{ijt}, \hat{\beta}_1), m_0(Z_{ijt}, \hat{\beta}_0)$). Doing so would mean estimating the gravity equation separately on EMU/non-EMU observations and then weighting the difference of conditional means of the outcome.[9]

In the related literature on free trade agreements, Baier and Bergstrand (2007) discuss the difficulties of estimating the causal effects using instrumental variable and control function approaches in panel data, suggesting that using panel fixed effects methods analogous to Eq. (1) should address many of the endogeneity concerns plaguing this literature. In the context of the euro effect on trade, the wide range of estimates documented in meta analysis such as Rose (2017) and Polák (2019) using similar panel fixed effects models suggests there are likely unaddressed endogeneity concerns here. One contribution of my work is to introduce the "doubly-robust" estimator in Eq. (3), which have been widely used elsewhere, to the trade literature. This provides a means of combining the potential benefits of the theory-consistent gravity approach[10] with information gained from estimates of selection into the EMU itself. Indeed, in later work, Baier and Bergstrand (2009) use traditional matching estimators for the effect of regional trade agreements. This method could combine such estimates with the gravity equation approach in Baier and Bergstrand (2007) to protect against mis-specification of either model. If the panel fixed effects model is correctly specified, then this approach should not affect results.

2.2 Modeling Selection Into Currency Unions

I estimate selection into currency unions using a logistic specification controlling for a wide range of country and pair-specific controls. I include in these a number of standard gravity equation variables, as well as other characteristics that may be important for the formation of bilateral currency union agreements. My first stage specification is given in Eq. (4).

$$\begin{aligned} EMU_{ijt} = \theta_0 \quad & + \theta_1 \ln Y_{it} \times Y_{jt} + \theta_2 \ln y_{it} \times y_{jt} + \theta_3 \ln Dist_{ijt} \\ & + \theta_4 \ln |Y_{it} - Y_{jt}| + \theta_5 \ln |y_{it} - y_{jt}| + \theta_6 \left| g_{it} - g_{jt} \right| \\ & + \theta_5 Population_{it} + \theta_6 Population_{jt} + \theta_7 Z_{ijt} + \epsilon_{ijt} \end{aligned} \tag{4}$$

In Eq. (4), in addition to the first three terms, which are standard gravity equation estimates of size of output ($Y_{it/jt}$), incomes ($y_{it/jt}$), and distances[11] ($Dist_{ijt}$) between the two countries, I also include the differences in output, per-capita GDP, and GDP

[8] Specifically I use the user-made reghdfe and ppmlhdfe packages in Stata with the inverse propensity score weights included as probability weights. For information on these procedures see: Guimarães and Portugal (2009), Guimarães and Portugal (2009), Correia et al. (2020), and Correia et al. (2019)

[9] This would simply allow the gravity coefficients other than EMU to be different across the two groups.

[10] Though any method of deriving an estimate of the conditional effect of conditional means for EMU/non-EMU in Eq. (3) could replace the gravity equation here.

[11] I use the population weighted distance between the two most populous cities in each country in thousands of kilometers.

growth rates ($g_{i/j}$). These are important because the standard gravity terms may do a poor job capturing the fact that many trading partnerships occur between relatively rich and poor (or large and small) economies, in ways that may be systematically different for the average eurozone economy. Further, I will include in my baseline specification a number of additional controls, including population of origin and destination countries, as well as a rich set of pair specific binary characteristics commonly used in the gravity literature[12] captured by Z_{ijt}. These controls may be important for capturing the various geopolitical motivations for forming a currency union. Moreover, Brookhart et al. (2006) shows that inclusion of variables related to the outcome of interest, even if unrelated to first stage treatment, decrease the variance of the estimated first stage without increasing bias. Further, Millimet and Tchernis (2009) show via Monte Carlo simulation that there are potential benefits of *over-specifying* the propensity score estimator, so I include squared terms of each of the continuous regressors in Eq. (4) along with their linear terms.

The first stage specification outlined above allows both weighting and sample truncation, but remains somewhat *ad hoc*. My emphasis, as I will show below, is on improving the comparability between treated and control observations in a way that is data driven, and that also easily links to the existing estimates on the euro trade effect. While I limit myself to factors commonly observed in trade data, other models using more detailed political/historical/demographic data may provide a better fit for the first stage selection process and improve upon the validity of my estimators. In part I wish to demonstrate, using widely available trade variables how differences already documented in Rose (2017) are closely related to the comparability of control groups.

3 First Stage Estimation and Sample Selection

To generate average treatment effects of the eurozone given by the IPWRA estimator of Eq. (3), I must first specify the probability of selection into the EMU given by a first stage estimation of Eq. (4). I report first stage estimates for a number of models, beginning first with one that simply uses traditional gravity equation estimates, then a specification that uses all of the controls described in the discussion of Eq. (4) excluding second order terms, then adding these second order controls to estimate my baseline specification. I consider two extensions of this baseline, first adding regional trade agreements and GATT membership at origin an destination, then adding to this European Union membership. The results of this first stage estimation is reported in Table 1.

The model generally finds sensible coefficients with EMU members having positive coefficients for log product of GDP and GDP per-capita, suggesting they have larger output and are richer than the average pair in the sample, but also negative

[12] In my baseline specification these are: whether the countries share a border, common language (both official and in terms of usage), common religion, common colonizers, and shared colonial histories.

Table 1 First Stage Models

	(1)	(2)	(3)	(4)	(5)
$log(GDP_o \times GDP_d)$	0.12***	0.69***	2.25***	0.69**	1.25**
	(0.01)	(0.01)	(0.25)	(0.26)	(0.45)
$log(GDPpc_o \times GDPpc_d)$	1.12***		12.80***	11.86***	11.70***
	(0.02)		(0.38)	(0.38)	(0.56)
$log\|GDP_o - GDP_d\|$		0.44***	-2.24***	-2.96***	-1.27***
		(0.02)	(0.18)	(0.19)	(0.35)
$log\|GDPpc_o - GDPpc_d\|$		-0.18***	-0.13***	-0.15***	-0.14**
		(0.01)	(0.03)	(0.03)	(0.04)
$\|Growth_o - Growth_d\|$		-15.33***	-15.56***	-15.36***	-16.07***
		(0.52)	(0.57)	(0.59)	(1.45)
Population Origin		-0.42***	0.00	-0.03	0.24***
		(0.01)	(0.02)	(0.02)	(0.07)
Population Destination		-0.40***	0.00	-0.03	0.28***
		(0.01)	(0.02)	(0.02)	(0.07)
Dist: (largest city, pop weighted)	-7.13***	-6.13***	19.01***	15.59***	10.77***
	(0.15)	(0.14)	(0.92)	(0.95)	(1.44)
Shared Border	-0.24***	-0.37***	0.65***	0.57***	0.75***
	(0.07)	(0.06)	(0.08)	(0.08)	(0.11)
Common Language (official)	0.35**	0.88***	1.82***	1.66***	-0.19
	(0.13)	(0.11)	(0.15)	(0.15)	(0.19)
Common Language (> 9%*pop.*)	-1.69***	-2.32***	-2.85***	-2.54***	0.82***
	(0.13)	(0.12)	(0.16)	(0.16)	(0.23)
Common Colonizer	-1.31***	-0.52**	-1.26***	-0.49*	1.48***
	(0.18)	(0.18)	(0.21)	(0.20)	(0.35)
Pair ever in colonial relationship	-0.74***	-0.52***	-0.01	-0.18	-1.08***
	(0.13)	(0.12)	(0.13)	(0.13)	(0.16)
Index of common religion	1.34***	1.79***	1.55***	1.61***	2.80***
	(0.06)	(0.06)	(0.07)	(0.07)	(0.10)
Origin GATT membership				0.73***	-0.33**
				(0.08)	(0.10)
Destination GATT membership				0.73***	-0.13
				(0.08)	(0.10)
Squared Continuous Vars			✓	✓	✓
RTA Control (dropped)				✓	✓
EU Membership Control (dropped)					✓
Observations	1,513,764	1,473,096	1,473,096	116,278	14,744
Pseudo R^2	0.619	0.595	0.698	0.539	0.405

Table 1 (continued)

	(1)	(2)	(3)	(4)	(5)
p-score range ($EMU = 1$)	[0.006,0.900]	[0.0001,0.853]	[0.001,0.893]	[0.004,0.965]	[0.011,0.995]
p-score mean ($EMU = 1$)	0.279	0.276	0.368	0.385	0.609
Omitted Obs. (due to RTA/EU)	0	0	0	1,356,818	1,458,352
p-score = 0 & trade != 0	0	3	142,912	3,557	3,717
0 < p-score < min p-sore($EMU = 1$)	699,624	745,085	705,265	86,294	0

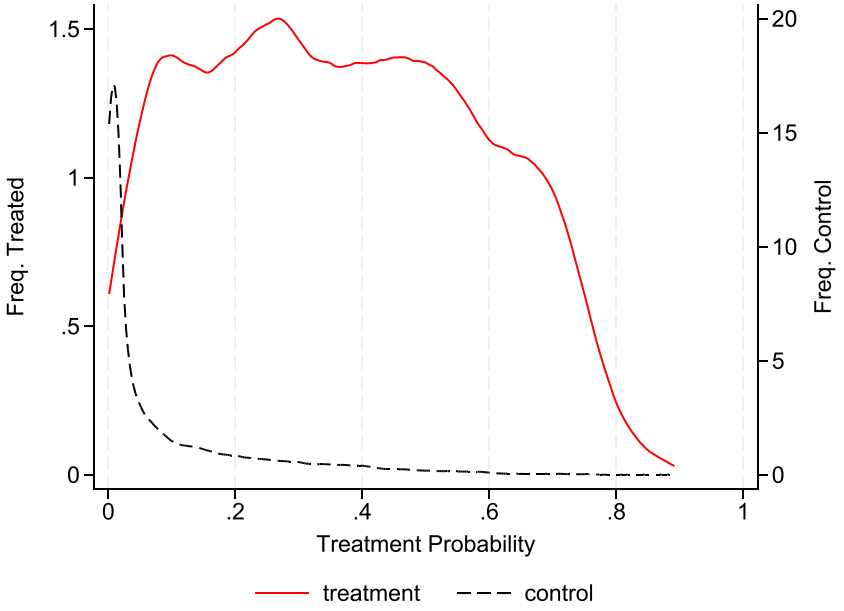
All estimations are logistic regressions with EMU membership as the outcome variable. Subscripts o/d refer to origin and destination respectively. Population in tens of millions and kilometers in ten thousands for readability. Standard errors reported in parenthesis. Regional trade agreement and EU membership dummies perfectly predict failure (ie all EMU members are in EU/RTAs with each other) and thus are dropped along with perfectly predicted observations. These can be included (using the "asis" option in Stata), but this creates instability in the likelihood function and such observations are assigned zero probability, dropping them from any p-score weighting

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$

terms for their absolute differences, which suggests they are closer in relative economic size and well being than the average trading partners. In Table 1 I exclude the non-linear squared terms present in models 3-5 to conserve space. The extended models (in columns 4 and 5) drop the vast majority of data as RTA and EU membership are both perfect predictors of "failure" (ie all EMU members have this attribute). Thus inclusion of one, or both, drops all zero observations for this variable.

Because fit is improved and Millimet and Tchernis (2009) suggests that IPWRA estimators may perform better with an *over-specified* model I will choose to use probabilities estimated using both the linear and squared terms of the variables in Eq. (4). Since I wish to compare my estimators to the log gravity estimates of Rose (2017), I keep baseline weighted estimate that is relatively close in sample size and will thus use the model from column 3 in Table 1 for my baseline estimates below. As I will show below, the estimates on my sample with non-zero predicted propensity scores in this model are quite similar to those in Rose (2017), making for a convenient baseline comparison for the rest of my results.

With such large data, and euro membership quite rare, the model struggles to fit this first stage particularly well, with the mean EMU member still having probability of joining at just 36.8%. A flaw of relying entirely on the strength of the matching estimator, as in work such as Persson (2001) and Chintrakarn (2008) is that these models have weaker fit, and less theoretical justification, than the gravity equation. An important strength of the IPWRA estimator is that it uses information from this first stage to rebalance treatment and control observations, while also relying on the gravity equations in 1 or 2 to calculate the conditional means across those groups. The first stage estimates in columns 4 and 5 do a better job identifying the factors that determine EMU membership, much better in the specification that drops non EU members. The inclusion of European Union membership as a control increasing the mean predicted probability of EMU members to 0.61. While I will not use these estimates in my baseline specification, as they are difficult to compare directly to



Graph shows kernel density plots of treatment and control sub-populations with propensity scores >= the minimum for the treated group.

Fig. 1 Overlap: K-density plots of treated and control propensity scores

work such as Glick and Rose (2016) that uses larger datasets, I will show that such estimation may improve comparability, but that estimates are similar ot my preferred specification using the baseline first-stage.

It is important that there is sufficient overlap in propensity score estimates, such that there are comparison observations across treatment and control groups. While the majority of weight of control population is near zero, I show in Fig. 1 that there is overlap across the full distribution of treated observations in my baseline (column 3) first stage model. Although I use separate axes for readability, there are more non-EMU observations with a probability greater than the mean of 0.368 than EMU observations.

Truncating the sample to only include observations along the [0.001, 0.893] distribution of the EMU observations results in a sample of 34, 971 (there are 4,690 pair-year observations between euro members). Such truncation is commonly used to limit the impacts of outliers. As can be seen in Eq. (3), EMU observations that are predicted to be very likely to be non-EMU and non-EMU observations with high predicted probability of being in the EMU can receive high weights at the extreme tails of the probability distribution. Many estimates making use of IPWRA and other propensity score estimates consider truncating the sample to limit such influence. Considering the weighting in Eq. (3) the threat is that very low probability treated observations (those that look much more like controls) and very

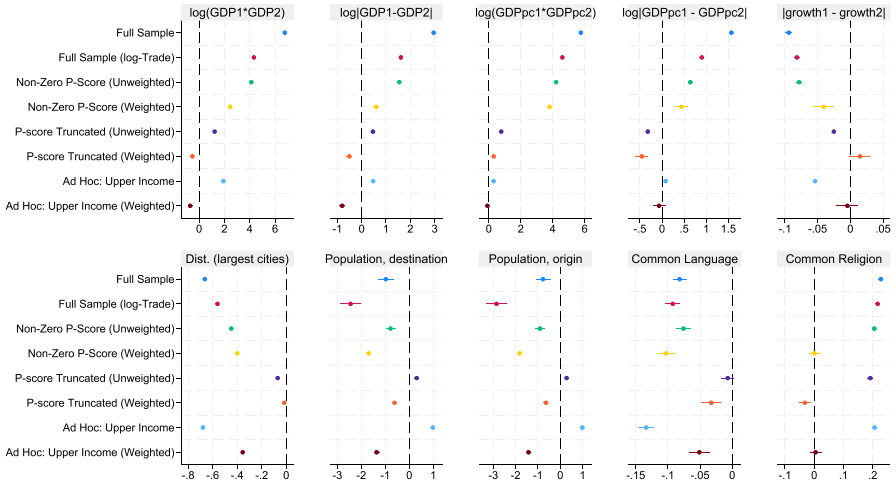


Fig. 2 Difference in Means (Treatment - Control), Various Controls

high probability controls (which look like treated) will receive potentially large weights. This often done using an arbitrary rule-of-thumb method such as 1% or 5% cutoffs. Imbens (2004) suggests that the potential threat of outliers shrinks with sample size so that the potential influence of low/high probabilities, providing $1/\left[N \times (1-\hat{p}_{max})\right]$ and $1/\left[N \times (\hat{p}_{min})\right]$ as bounds for the influence of observations at the top and bottom of the probability distribution. Using the sample that lies along the range of the treatment observations ([0.001, 0.893]) limits the potential effect of outlier treatment and control impact on the estimator to less than 3%, well within the bounds used in other studies.

More importantly in this context and for the discussion of proper choice of sample in Rose (2017), it would appear in large trade datasets that the *opposite* problem to that of influential outliers may be problematic for the Eq. (3). Rather than being driven by small(large) treatment(control) outliers, estimates on the full sample are strongly influenced by the large number of Non-EMU observations with extremely low probability of treatment. These receive smaller weights in Eq. (3), but the weighting penalty only slightly diminishes this group's out-sized effect on the estimators, as I will show. This problem is larger in unweighted estimates of Eqs. (1) and (2). Another way this can be shown is by *censoring* p-scores (keeping the observation but setting an upper/lower threshold for their weights), which has little effect relative to IPWRA estimates on the full data. The fact that truncating these observations changes their results dramatically, while assigning arbitrary lower/upper probability limits on the full sample does not, implies that the large weight of low probability observations are altering estimates in spite of their low weights, not because of them. For clarity of exposition I do not include such censored results below.

I agree with Rose (2017) that sample choice is critical in the estimation of EMU trade effects. Before presenting estimates of the EMU I show that *how* the sample is selected points towards preferring estimates made on a smaller subset of data.

I consider five samples, built in various ways using the propensity scores from column 3 of Table 1. These are: the full sample including zero and missing trade flows,[13] the full sample with non-zero trade (ie using the log specification), that for which predicted probabilities of treatment are non-zero (ie the largest sample where the IPWRA is estimable), a truncated sample with probabilities limited to the range of the treatment group, and finally an *ad hoc* sample of upper income countries. This last group is consistent with the definition in Rose (2017), and includes only origin and destination countries with real GDP per-capita greater than $12,736.

Figure 2 shows the difference in means between EMU (treatment) and non-EMU (control) observations across each of these samples for many of the controls used in Eq. (4). I provide both the unweighted and weighted differences in means using propensity score estimates when possible. In an ideal situation, such as a properly run randomized control trial, these groups should be identical along observable characteristics and these differences should be zero. This exercise highlights the value that the first stage propensity score process has in improving the comparability of these two groups.

Figure 2 shows that the set of EMU countries are not remotely comparable to the mean control in the full sample. The first and third graph suggest that eurozone partner GDPs and GDP per-capita are substantially larger than the mean in the sample. The absolute value differences in the second, fourth, and fifth sub-figures reflect that fact that EMU trading partners are more diverse in terms of relative output and income than the average pair, but on much closer growth trajectories. Unsurprisingly EMU members are much closer together geographically, they also have relatively smaller populations, are less likely to share a common (official) language, and much more likely to have a common religion.

The absolute difference in means for the full sample is nearly always largest,[14] with the non-missing trade already improving the comparability between the two groups. However in the full log sample, which is smaller, but of a similar magnitude and composition to that used in Glick and Rose (2016), differences are still extremely far from zero. Because these differences are precisely estimated, many of the error bands are not visible, but they are strongly significant. The improvement when moving from full to non-zero/missing trade is somewhat intuitive given that these flows are concentrated in small economies. Comparing the largest p-score samples, weighting improves comparability, but in some cases only marginally. Truncating the sample to include only the non-EMU observations that fall within the range of estimated p-scores of EMU pairs brings these much closer to zero, with weighting further closing the remaining gap for all but one (difference in GDP per capita) control. Though these small differences are in some cases still significant they provide a much stronger comparability between the EMU sample and the comparison group used in estimation.

It is interesting to compare these model driven sample means with the *ad hoc*, "upper income" sample used in Rose (2017). For the control used to restrict the

[13] For PPML cases I assume missing trade data reflects a zero trade flow in cases where the GDP data for origin and destination countries is non-missing, leaving remaining cases as missing.

[14] The notable exception is in population, where missing trade flows are strongly correlated with smaller countries, whose inclusion brings the sample average closer to the relatively smaller euro average.

sample (GDP per capita) this actually outperforms the comparability made using propensity scores. The model in Eq. (4) works to minimize gaps along all of these variables, many of which substantially outperform the *ad hoc* measure, at the expense of fitting the per-capita GDP comparisons as closely. The final sample, which combines the *ad hoc* weights with the logistic model propensity scores does improve the fit (though not generally as well as the p-score trucation), but I caution that this actually further truncates the sample as a non-trivial share of the upper-income countries were allocated scores of zero and thus drop out when the inverse-propensity score weights are included. Authors wishing to use such *ad hoc* sample reduction techniques should check that it improves the balance of treated/control observations, and consider whether a data driven approach might futher improve them.

One can consider a reduction in sample using IPWRA estimates as a data-driven approach to the sample selection, providing the best possible comparability between treated and control, conditional on observable characteristics, while also improving comparability through the weighting procedure itself. Of course, such attempts at constructing a re-randomized treatment and control set ex-post are only as reliable as the observable characteristics used in the first-stage selection model. It is still possible that unobserved characteristics not captured in this process may bias estimates. In the *doubly robust* estimator used in Eq. (3), or estimates that rely on the propensity score model to only truncate the sample, this only creates a problem if they are *also* not captured by the rich set of origin-year, destination-year, and dyadic fixed effects used in the second stage gravity equation. While this is still potentially true, my results should improve upon the existing estimates of the EMU treatment effect in the aggregate trade data by providing a nearly balanced treatment and control group prior to estimating the marginal effect using conventional gravity measures.

4 Results: IPWRA Gravity Estimates of Trade

I now present results for Eqs. (1) and (2) on the samples described above, showing when possible the improvements made by weighting as in Eq. (3). Because I wish to understand which changes come from weighting and which come from sample selection, I will include the OLS/PPML without the IPWRA estimate for each of these subsamples. I begin with the OLS estimates of the log gravity specification in Table (2).

The estimate using the full sample is larger, at 0.48 than the 0.43 coefficient from Glick and Rose (2016).[15] This effect is essentially unchanged when running the same specification on the smaller sample that omits propensity scores assigned an exact zero by the first stage model.[16] This is perhaps not too surprising given

[15] In what follows I report coefficient estimates, not the percent change in trade. For my lower truncated results these are quite close to each other. My starting coefficient of 0.48 is an EMU effect that implies a 59.5% ($e^{0.48} - 1$) increase in bilateral trade as a result of the euro.

[16] Due to issues of precision this is actually slightly smaller as some non-zero (but extremely low) propensity scores get assigned a zero when constructing the inverse weights in Stata.

Table 2 Log Gravity Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
EMU	0.483***	0.475***	0.404***	0.076***	0.048**	0.133***	0.025
	(0.017)	(0.018)	(0.018)	(0.024)	(0.024)	(0.028)	(0.028)
RTA	0.323***	0.206***	0.203***	0.007	0.008	0.059*	0.066*
	(0.008)	(0.009)	(0.009)	(0.027)	(0.027)	(0.030)	(0.034)
Ex-Year FEs	✓	✓	✓	✓	✓	✓	✓
IM-Year FEs	✓	✓	✓	✓	✓	✓	✓
Dyadic FEs	✓	✓	✓	✓	✓	✓	✓
Sample:							
p-score		✓	✓				
truncated				✓	✓		
upper-income						✓	✓
IPWRA			✓		✓		✓
R2	0.871	0.883	0.891	0.968	0.973	0.949	0.964
N	810,018	558,226	558,226	34,971	34,971	75,311	45,391

All models report OLS estimates of log bilateral export flows as the dependent variable. Heteroskedasticity-robust standard errors reported in parenthesis. IPWRA models use propensity score weights in regression as described in Eq. (3). P-score sample includes all observations with non-missing or zero first stage probability of treatment, truncated sample drops all observations outside the range of first stage probability among EMU country-pairs, while upper-income restricts the sample to countries with per capita GDP below 12,736 as in Rose (2017)

$p < 0.15$; $*p < 0.10$; $**p < 0.05$; $***p < 0.01$

that these samples have fairly similar differences in means across treatment and control in Fig. 2. When applying the IPWRA estimator to weight the regression outcomes by probability of selection this estimate is reduced to 0.40, but remains large and quite close to the Glick and Rose (2016) estimate for the EMU. Estimates on the smaller, truncated, sample using only observations within the range of the propensity scores observed for the EMU, (in the range [0.001, 0.893]) decrease these effects substantially to 0.076. On the truncated sample, the weighted coefficient is close to, but roughly a standard error below, the unweighted estimate. The propensity score is still important for the unweighted sample given that it was used to select inclusion, but this selection appears to matter much more than the weighting. For the *ad hoc* Upper income sample, I estimate a 0.13 EMU effect, quite similar to the equivalent estimate of 0.11 in Rose (2017). Estimates using this sample selection are nearly twice as large as those on the truncated propensity score sample. Adding the propensity score weights to this sample, to estimate the IPWRA treatment effect, reduces this effect to be indistinguishable from zero, though notably there is a large fraction of the this sub-sample that have zero estimated propensity scores, and therefore drop out of this estimation. Repeating the estimation using the sample in column 7 without weighting provides an estimate of 0.03, so as with the truncated sample most of this difference appears to come from the selection effect of dropping observations with extremely poor fit in the first stage model, rather than the weighting estimator itself.

Table 3 PPML Gravity Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
EMU	0.085***	0.069***	0.059***	0.071***	0.050**	0.006	0.000
	(0.011)	(0.011)	(0.011)	(0.025)	(0.024)	(0.014)	(0.015)
RTA	0.057***	0.144***	0.145***	0.076*	0.071*	0.070***	0.125***
	(0.010)	(0.007)	(0.007)	(0.038)	(0.039)	(0.014)	(0.015)
Ex-Year FEs	✓	✓	✓	✓	✓	✓	✓
IM-Year FEs	✓	✓	✓	✓	✓	✓	✓
Dyadic FEs	✓	✓	✓	✓	✓	✓	✓
Sample:							
p-score		✓	✓				
truncated				✓	✓		
upper-income						✓	✓
IPWRA			✓		✓		✓
pseudo-R2	0.991	0.991	0.993	0.997	0.997	0.996	0.996
N	1,521,887	558,226	558,226	34,971	34,971	86,971	45,391

All estimates use Poisson pseudo-maximum likelihood estimators with bilateral export flows as the dependent variable. Heteroskedasticity-robust standard errors reported in parenthesis. IPWRA models use propensity score weights in regression as described in Eq. (3). P-score sample includes all observations with non-missing or zero first stage probability of treatment, truncated sample drops all observations outside the range of first stage probability among EMU country-pairs, while upper-income restricts the sample to countries with per capita GDP below 12,736 as in Rose (2017)

$p < 0.15$; $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Repeating this exercise for the PPML estimates of the gravity equation provide substantially different results, and are shown in Table 3. The first column uses the full sample, assuming zero trade flows for missing values where the macroeconomic controls (GDP and GDP per capita) are non-missing.[17] The resulting estimate of the EMU effect is 0.085, larger than the equivalent estimate of 0.052 in Larch et al. (2019), which again I attribute to using a different sample and not jointly controlling for other currency unions (rather dropping their observations). Now when reducing the sample to those with positive propensity scores this falls only slightly to 0.069, and again to 0.05 with the truncated propensity score sample. These estimates for the PPML are all quite close to those from the log-gravity estimation using the truncated sample.

While my preferred truncated sample IPWRA estimates remain robust in this PPML estimation using heteroskedastic-robust standard errors, Egger and Tarlea (2015) suggest using multi-way exporter, importer, and year clustering in gravity equation estimations. Consistent with the PPML results of Larch et al. (2019), who

[17] I note that while I include missing flows as zero in column 1, column 2 is the same sample as in Table 2 as these observations get assigned missing or zero propensity scores. The large differences in these initial estimates come almost entirely from the PPML model and not from inclusion of these missing flows.

cite both Huber/White and multi-way clustered errors of this type, the significance for all of my results using the PPML estimator fail to reach even a $p < 0.15$ threshold when implementing these standard errors. I report these more forgiving standard errors here as they are those used in Glick and Rose (2016) and Rose (2017), with which I wish to draw closest comparison. I caution that in addition to all PPML estimates, the log gravity estimates on the truncated sample also lose statistical significance using exporter-importer-year multi-way clustering.[18]

An interesting result in Table 3 is the consistency of estimates across samples in columns 2-5. Once dropping the observations with a zero estimated probability of treatment the PPML estimator appears quite robust to sample selection, with all estimates extremely close to my preferred specification using the log-gravity specification in column 5 of Table 2. It is also notable that the *ad hoc* sample, which in general was a poorer fit than the truncated p-score sample in Fig. 2, has lower estimates. I interpret the results in this table as suggesting that the PPML estimate is considerably more robust to the selection issues from log-gravity specifications. While the full and *ad hoc* samples differ dramatically, the large reduction in sample when using the full propensity score sample to the truncated sample results in only small differences in my estimates. The differences when using standard PPML and the IPWRA framework are also small. Researchers truncating in arbitrary ways should be careful to at least justify such choices based on comparability of treatment and control and consider if they might be throwing away good control observations (while including some bad ones), as this exercise might suggest.

4.1 Robustness to First Stage Specification

Above I present baseline results using the third model in Table 1. This is my preferred specification for the paper for two primary reasons. The first is that it maximizes first stage fit, as measured by pseudo-R^2. The second, and importantly for my motivation for study, is that it retains enough data in the non-truncated sample to demonstrate that my log gravity results closely mirror the large EMU effects found in work such as Glick and Rose (2016).

A logical question would be whether or not the above results are highly sensitive to the choices made in this first stage model. To answer this I replicate my main propensity score results for each of the models in Table 1 to show how the implied trade effect differs across them. These are reported in Table 4 in the same order they are presented in Table 1 (the baseline estimates are thus the third and included for ease of comparison). To conserve space, I exclude the full sample estimates (which are by definition unchanged) as well as the *ad hoc* upper income sample selection. For the latter the unweighted sample estimates would be identical across specifications and the weighted estimates in column 7 of Tables 2 and 3 differ both due to weighting and due to further sample selection, making the source of variation difficult to compare across these groups.

[18] Though the results in columns 1-4 of Table 2 remain significant (at varying levels) when using more conventional exporter or pairwise clustering.

Table 4 EMU Effect on Trade: First Stage Robustness

	(1) log	(2) log	(3) log	(4) log	(5) PPML	(6) PPML	(7) PPML	(8) PPML
	Simplified First Stage: Traditional Gravity Controls							
EMU	0.519***	0.437***	0.031	0.034	0.074***	0.071***	0.032	0.052**
	(0.017)	(0.017)	(0.022)	(0.022)	(0.011)	(0.011)	(0.021)	(0.022)
R2	0.876	0.887	0.959	0.969	0.991	0.993	0.996	0.997
N	701,581	701,581	47,915	47,915	701,581	701,581	47,915	47,915
	Baseline First Stage: Linear Controls Only							
EMU	0.519***	0.426***	0.366***	0.283***	0.074***	0.068***	0.074***	0.071***
	(0.017)	(0.018)	(0.020)	(0.020)	(0.011)	(0.011)	(0.014)	(0.015)
R2	0.876	0.890	0.913	0.928	0.991	0.993	0.992	0.995
N	693,455	693,455	223,673	223,673	693,455	693,455	223,673	223,673
	Baseline First Stage							
EMU	0.476***	0.405***	0.075***	0.048*	0.069***	0.059***	0.071***	0.050**
	(0.018)	(0.018)	(0.024)	(0.024)	(0.011)	(0.011)	(0.025)	(0.024)
R2	0.883	0.891	0.968	0.973	0.991	0.993	0.997	0.997
N	558,157	558,157	34,961	34,961	558,157	558,157	34,961	34,961
	Baseline First Stage: Including Trade Agreements Membership							
EMU	0.072***	0.041*	0.031	0.010	0.045**	0.034	0.071**	0.052*
	(0.022)	(0.022)	(0.022)	(0.022)	(0.021)	(0.021)	(0.027)	(0.026)
R2	0.955	0.961	0.981	0.984	0.997	0.997	0.996	0.997
N	98,637	98,637	23,035	23,035	98,637	98,637	23,035	23,035
	Baseline First Stage: Including European Union Membership							
EMU	-0.032	-0.034	0.025	0.007	-0.030*	0.009	0.036**	0.068***
	(0.021)	(0.025)	(0.026)	(0.030)	(0.016)	(0.017)	(0.017)	(0.018)
R2	0.988	0.989	0.990	0.991	0.998	0.998	0.998	0.998
N	14,738	14,738	11,502	11,502	14,738	14,738	11,502	11,502
Ex-Year FEs	✓	✓	✓	✓	✓	✓	✓	✓
IM-Year FEs	✓	✓	✓	✓	✓	✓	✓	✓
Dyadic FEs	✓	✓	✓	✓	✓	✓	✓	✓
	Sample							
p-score	✓	✓			✓	✓		
truncated			✓	✓			✓	✓
IPWRA		✓		✓		✓		✓

Log estimates use OLS estimates of log bilateral export flows, while PPML use the Poisson pseudo-maximum likelihood estimator on bilateral export flows. Heteroskedasticity-robust standard errors reported in parenthesis. P-score sample refers to sample with non-zero probability of treatment, while truncated refers to sample reduction based on the probability range of treated (EMU) observations. IPWRA refers to regression adjustment through propensity score weighting as described in Eq. (3).

$p < 0.15$; $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

The first result to note in Table 4 is that log gravity estimates are much more sensitive to these first stage specifications than PPML estimates. This exercise should serve as a strong advertisement for the robustness of the PPML estimator. The

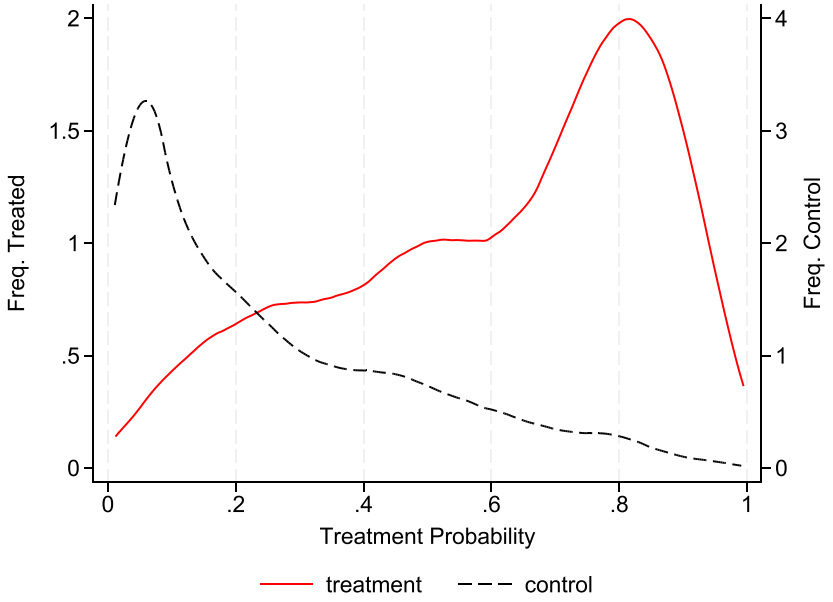


Fig. 3 Overlap: K-density plots of treated and control propensity scores (EU first stage)

second specification, using the same set of variables as my baseline but with only linear controls, keeps much more data on truncation than my other specifications. This is because the model fits much lower values for some euro members, and I keep the criteria described above for truncating based on the support of the treated observations. Since the minimum of p-score for this model is 0.0001, this a large amount of relatively poorly fitting controls are now retained. As a result these are the only results on the truncated sample that differ substantially from the baseline. The preferred truncated estimates in the PPML on the other hand are quite similar across all models, with weighting seeming to consistently pull the point-estimate in the direction of the 5% EMU trade effect found in the baseline.

I report the overlap and differences in means for treated and control observations for each of these in Appendix A. However, it is worth devoting some space to exploring these for the model that includes European Union membership. Recall from Table 1 that while this model had lower fit, as measured by pseudo-R^2, it had by far the highest mean probability for the treated sub-population. In Fig. 3 I report the k-densities of the estimated probability of treatment across the treatment and control groups using this sample. The plots are for the range of propensity scores among the EMU population, which for this specification is: [0.011, 0.995]. This is

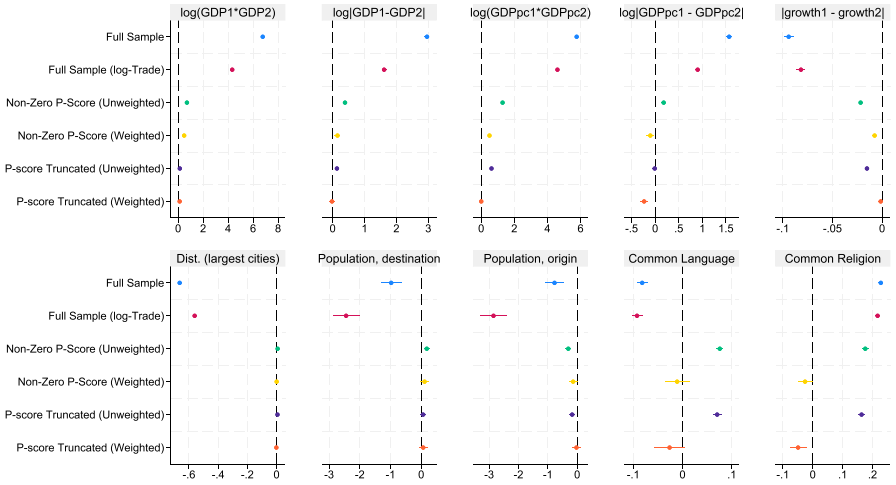


Fig. 4 Difference in Means (Treatment - Control), EU First Stage

an ideal first-stage overlap figure as it demonstrates both overlap across both groups, but also is clearly capable of discerning EMU membership from non-members. In many ways this addresses the main concern that Rose (2001) made in response to the Persson (2001) matching estimator.

Turning to the differences in means, it is clear in Fig. 4 that this model brings the differences between these two groups quite close to zero, particularly with weighting and truncation, where only two of the ten controls shown have means that are significantly different from each other at the 5% level. Part of the reason for an improved fit, is that this sample is limited to EU countries with non EMU members of the broader union sharing many characteristics with their eurozone counterparts. In some ways this model can be seen of a hybrid of simpler *ad hoc* measures with my data driven approach. By including a perfect predictor, like EU membership, the sample becomes highly constrained, but in a way that provides actual benefits in terms of improving the comparability between the treated and control groups. The propensity score model still works to improve this fit in two ways. First through weighting, as can be seen by the shifts toward zero from the unweighted versions (non-zero and truncated) when propensity score weights are added. The second way is by removing poor comparisons through truncation, demonstrated in the shift toward zero from the non-zero p-score (unweighted) estimates to the p-score truncated (unweighted) group.

5 Conclusions

Eurozone membership was not assigned by a researcher, but the culmination of intense policy deliberation. While it may not be possible to fully rectify selection effects of this policy's effect on trade, propensity score methods offer a way of improving the credibility of estimators to take such differences between EMU and

non-EMU pairs into account. Log gravity equation estimates of the EMU treatment effect are extremely sensitive to sample changes. I argue that this sensitivity largely reflects improvements in the control sample, bringing their observable characteristics closer to those of EMU countries. Ad hoc measures do a good job fitting on a given statistic, as my example using a GDP per capita cutoff to determine selection into the estimation sample, but cannot balance the potential trade-off between improving comparability along many dimensions. I argue that data driven propensity score approaches provide a better way of improving the comparability of these groups, while not relying on potentially arbitrary decision making of researchers. Use of the doubly robust IPWRA estimator provides a method for re-weighting the average treatment effects from gravity equations to further improve this fit, but in practice has small quantitative implications relative to the effects of truncation.

Results using samples where observable EMU and non-EMU characteristics are most comparable suggest a small and positive impact of the currency union on trade. This is roughly a 5% increase that is quite similar across my preferred log and PPML specifications. Limiting the sample to estimate thse effect among European Union measures might increase this slightly to nearly 7%. These are in line with the survey of the literature by Polák (2019), and explain why the large estimates of Glick and Rose (2016) are likely biased upwards. While Rose (2017) identifies the correct reason for differences in trade estimates of the euro, the evidence presented here suggests that his conclusion that more data should be better, is likely mistaken.

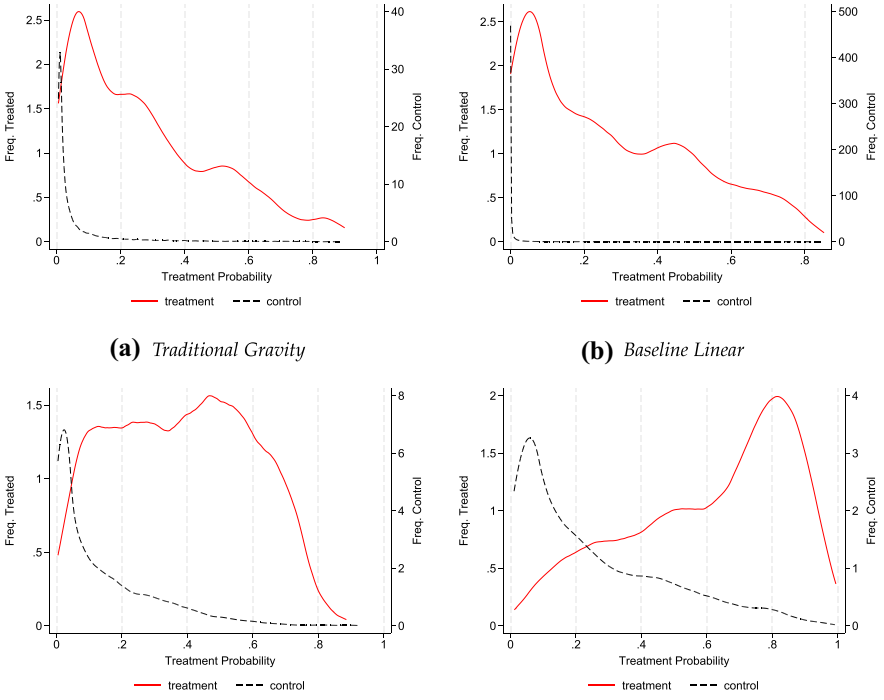
More broadly my results suggest that greater care should be made in sample selection in trade and other macro-policy environments. Observational studies can leverage models of first stage treatment to ensure better balance between countries that are exposed to a particular policy and those that are not. While this does not completely bridge the gap in causal identification between observational studies and the experimental ideal, it provides an empirically driven step in the right direction. The PPML estimator is much more robust to changing sample, though still reflects statistically meaningful shifts when moving from the full sample to those that drop the most extreme outliers and when using the IPWRA adjustment. This suggests that research using log approximations should be particularly careful of these sample selection issues.

In my preferred estimates the EMU effect is small, but still meaningful. Future work might seek to better model the selection process, complementing trade data with richer controls to better capture the geopolitical decision making process. Theoretical grounding in the literature on optimum currency areas stemming from Mundell (1961) may also prove a fruitful avenue for improving on the estimates presented here. Empirically these methods may be useful in quantifying the role that the euro has played on trade between eurozone countries and other partners, as Martínez-Zarzoso (2019) explores for EMU trade with the CFA franc countries. Another potential application is exchange regime choice and currency networks, as studied in Bleaney and Tian (2012), to understand the role that such selection has on estimates of exchange rate regimes on other macroeconomic outcomes. Better understanding the such additional channels through which membership in the EMU affects finance and trade will be important in getting a full picture of the economic benefits enjoyed by member states.

Appendix A: Performance of Alternative First-stage Models: Overlap and Comparability

In this section I present information on overlap and difference in means for all of the non-baseline models from Table 1. These are equivalent to the k-density graph from Fig. 1 and the comparability of treatment and control groups (for select variables) as shown for the baseline case in Fig. 2. These are shown for each of the four alternative first-stage specifications from Table 1 which control for: traditional gravity variables (Table 1, Column 1), the baseline with only linear control terms (Table 1, Column 2), inclusion of regional trade agreements and GATT membership (Table 1, Column 4), and finally adding these trade agreements as well as European Union (EU) membership to the baseline (Table 1, Column 5).

Overlap is plotted in Fig. 5. Here the goal is a model that both does a good job predicting treatment (assigns high probability to treated observations) while still having enough overlap between treated and control observations. On this count, the broad takeaway is that the first two models, using traditional gravity controls only, or using my baseline model without additional quadratic controls do a significantly poorer job than the baseline model. Inclusion of regional trade agreements (which substantially reduces the sample) as well as

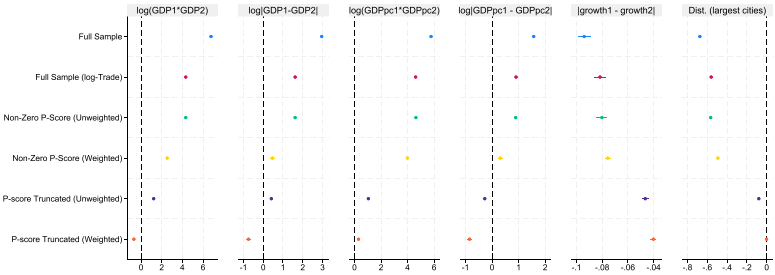


(a) *Traditional Gravity* **(b)** *Baseline Linear*

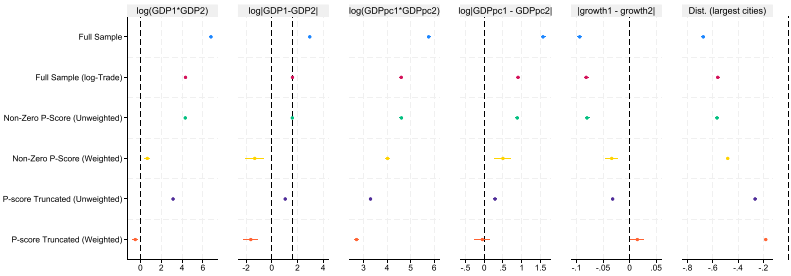
(c) *Baseline + Trade Agreements* **(d)** *Baseline + Trade Agreements + European Union*

Graph shows kernel density plots of treatment and control sub-populations with propensity scores >= the minimum for the treated group.

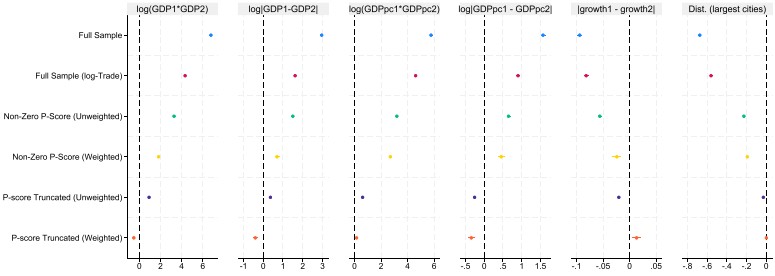
Fig. 5 Overlap: K-density plots of treated and control propensity scores: All Models



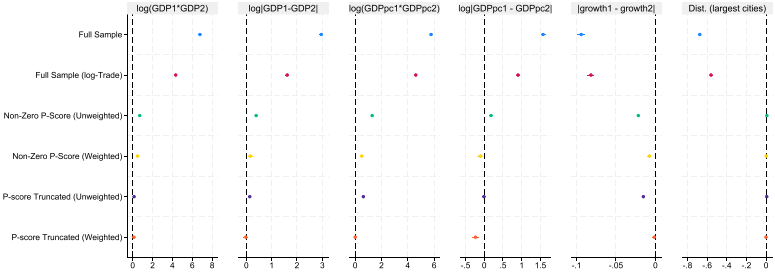
(a) *Traditional Gravity*



(b) *Baseline with Linear Terms*



(c) *Baseline + Trade Agreements*



(d) *Baseline + Trade Agreements + European Union*

Difference in Means for select variables using additional first stage weighting and truncation.

Fig. 6 Difference in Means: Treatment - Control

GATT membership looks quite similar to the baseline estimates presented in Fig. 1, and so potentially drops data without large gains in fit. Finally using European union in the first stage, which restricts the sample entirely to countries that are within the EU does a great job both in terms of discernment of treatment and overlap between treated and control.

Plotting differences in means for some of the main controls used in the estimation we see quite a similar pattern to what was reported in Fig. 6. Generally the more controls added the closer the final two rows, which reflect the truncated propensity score sample with and without weighting. Appear to perform the best, with the European Union sample bringing the differences in means to zero for all but one of these variables in the case of both truncation and weighting.

This evidence suggests that by the criteria of improving first stage predictability and differences in model means the first stage that includes regional trade agreements, GATT membership in both origin and destination, and European Union membership does by far the best job. One downside is the loss of data, making estimates hard to compare to the traditional estimates in Glick and Rose (2016), which is why this model wasn't chosen as a baseline, but I do not view the results of the baseline as superior to those that use the more selective first stage.

Appendix B: Controlling for European Union Membership

Table 5 Log Gravity Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
EMU	0.263***	0.291***	0.232***	0.074***	0.046*	0.095***	0.016
	(0.017)	(0.018)	(0.017)	(0.024)	(0.024)	(0.028)	(0.028)
RTA	0.300***	0.193***	0.189***	0.008	0.010	0.051*	0.068*
	(0.008)	(0.009)	(0.009)	(0.027)	(0.027)	(0.030)	(0.034)
EU	0.586***	0.589***	0.549***	0.042	0.056	0.382***	0.203***
	(0.014)	(0.016)	(0.015)	(0.038)	(0.036)	(0.056)	(0.054)
Ex-Year FEs	✓	✓	✓	✓	✓	✓	✓
IM-Year FEs	✓	✓	✓	✓	✓	✓	✓
Dyadic FEs	✓	✓	✓	✓	✓	✓	✓
Sample:							
p-score		✓	✓				
truncated				✓	✓		
upper-income						✓	✓
IPWRA			✓		✓		✓
R2	0.871	0.883	0.891	0.968	0.973	0.949	0.964
N	810,018	558,226	558,226	34,971	34,971	75,311	45,391

All models report OLS estimates of log bilateral export flows as the dependent variable. Heteroskedasticity-robust standard errors reported in parenthesis. IPWRA models use propensity score weights in regression as described in Eq. (3). P-score sample includes all observations with non-missing or zero first stage probability of treatment, truncated sample drops all observations outside the range of first stage probability among EMU country-pairs, while upper-income restricts the sample to countries with per capita GDP below 12,736 as in Rose (2017)

$p < 0.15$; *$p < 0.10$; **$p < 0.05$; ***$p < 0.01$

Table 6 PPML Gravity Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
EMU	0.080***	0.067***	0.058***	0.072***	0.051**	0.005	0.000
	(0.011)	(0.011)	(0.011)	(0.025)	(0.024)	(0.014)	(0.015)
RTA	0.052***	0.140***	0.141***	0.078**	0.073*	0.069***	0.125***
	(0.010)	(0.007)	(0.007)	(0.039)	(0.039)	(0.014)	(0.015)
EU	0.328***	0.237***	0.231***	0.060**	0.070**	0.068*	0.015
	(0.014)	(0.013)	(0.013)	(0.028)	(0.031)	(0.035)	(0.030)
Ex-Year FEs	✓	✓	✓	✓	✓	✓	✓
IM-Year FEs	✓	✓	✓	✓	✓	✓	✓
Dyadic FEs	✓	✓	✓	✓	✓	✓	✓
Sample:							
p-score		✓	✓				
truncated				✓	✓		
upper-income						✓	✓
IPWRA			✓		✓		✓
pseudo-R2	0.991	0.991	0.993	0.997	0.997	0.996	0.996
N	1,521,887	558,226	558,226	34,971	34,971	86,971	45,391

All estimates use Poisson pseudo-maximum likelihood estimators with bilateral export flows as the dependent variable. Heteroskedasticity-robust standard errors reported in parenthesis. IPWRA models use propensity score weights in regression as described in Eq. (3). P-score sample includes all observations with non-missing or zero first stage probability of treatment, truncated sample drops all observations outside the range of first stage probability among EMU country-pairs, while upper-income restricts the sample to countries with per capita GDP below 12,736 as in Rose (2017)

$p < 0.15$; $*p < 0.10$; $**p < 0.05$; $***p < 0.01$

Rather than controlling for European Union membership in the first stage as in Table 4, one could include this as a control. Recent work by Felbermayr et al. (2022) look at the various degrees of integration in Europe to quantify the potential costs of undoing European integration. While I do not fully replicate the various stages studied there (eg: Schengen area), a simple test of robustness of my trade estimates to the EMU is to include the European Union as an additional control. I present these for both the log gravity and PPML estimates in Tables 5 and 6. In both cases inclusion of a European Union dummy has almost no bearing in addition to the regional trade agreement variable used in the baseline estimates, suggesting that at least in broad terms my results for the EMU are not capturing some broader impact of European integration.

Future work might seek to more carefully estimate the dynamic impacts the EMU, as is studied in Bergin and Lin (2012). There they find significant anticipation effects, which may be relevant here, and such a framework would lend itself to studying the various stages of integration as studied in Felbermayr et al. (2022).

Appendix C: Changes in Choice of Truncation

In this appendix I experiment with changing the level of truncation. As the baseline estimates (reported here in Tables 7 and 8 columns 1 and 2) use the support of the treatment group, these drop any p-scores outside of the [0.001, 0.893] range. In practice this only drops variables below this as no treated observations are assigned probability above. Thus I experiment with moving the lower bound of this truncation restriction to be $0.001/5 = 0.0002$ and $0.001 \times 5 = 0.005$. These are arbitrary changes meant to investigate the sensitivity of estimates in the area around truncation, but allow for substantial increase/decreases in the sample around this estimate.

Generally speaking the effect of altering the sample in this way is small. For log estimates, the implication is similar to that of estimates in the paper, which suggest that more data increases both the unweighted and IPWRA estimates. This is not the case for PPML estimates, where coefficients are systematically lower in for both the increased and decreased sample. However, in both cases estimates remain within a standard error of those found in the baseline model. Moreover, the various models tested in Table 4 involve large changes to the truncation and suggest that the PPML estimates are strongly robust to such changes when determined by various first stage model fits.

Table 7 Log Gravity Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
EMU	0.076***	0.048**	0.094***	0.065**	0.063**	0.042*
	(0.024)	(0.024)	(0.024)	(0.023)	(0.025)	(0.024)
RTA	0.007	0.008	0.014	0.015	-0.006	0.000
	(0.027)	(0.027)	(0.026)	(0.025)	(0.030)	(0.031)
Ex-Year FEs	✓	✓	✓	✓	✓	✓
IM-Year FEs	✓	✓	✓	✓	✓	✓
Dyadic FEs	✓	✓	✓	✓	✓	✓
Sample:						
Truncated: Support of treatment	✓	✓				
Truncated: Min treatment/5			✓	✓		
Truncated: Min treatment× 5					✓	✓
IPWRA		✓		✓		✓
R2	0.968	0.973	0.963	0.969	0.976	0.980
N	34,971	34,971	44,733	44,733	25,546	25,546

All models report OLS estimates of log bilateral export flows as the dependent variable. Heteroskedasticity-robust standard errors reported in parenthesis. IPWRA models use propensity score weights in regression as described in Eq. (3)

$p < 0.15$; *$p < 0.10$; **$p < 0.05$; ***$p < 0.01$

Table 8 PPML Gravity Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
EMU	0.071***	0.050**	0.050**	0.034	0.063**	0.041
	(0.025)	(0.024)	(0.023)	(0.022)	(0.028)	(0.027)
RTA	0.076*	0.071*	0.035	0.030	-0.052*	-0.057**
	(0.038*)	(0.039)	(0.039)	(0.039)	(0.028)	(0.028)
Ex-Year FEs	✓	✓	✓	✓	✓	✓
IM-Year FEs	✓	✓	✓	✓	✓	✓
Dyadic FEs	✓	✓	✓	✓	✓	✓
Sample:						
Truncated: Support of treatment	✓	✓				
Truncated: Min treatment/5			✓	✓		
Truncated: Min treatment× 5					✓	✓
IPWRA		✓		✓		✓
pseudo-R2	0.997	0.997	0.997	0.997	0.996	0.997
N	34,971	34,971	44,733	44,733	25,546	25,546

All estimates use Poisson pseudo-maximum likelihood estimators with bilateral export flows as the dependent variable. Heteroskedasticity-robust standard errors reported in parenthesis. IPWRA models use propensity score weights in regression as described in Eq. (3)

$p < 0.15$; *$p < 0.10$; **$p < 0.05$; ***p < 0.01

Funding Open Access funding provided by the IReL Consortium.

References

Aizenman J (2018) Optimal currency area: a twentieth century idea for the twenty-first century? Open Econ Rev 29:373–382

Baier SL, Bergstrand JH (2007) Do free trade agreements actually increase members' international trade? J Int Econ 71(1):72–95

Baier SL, Bergstrand JH (2009) Estimating the effects of free trade agreements on international trade flows using matching econometrics. J Int Econ 77(1):63–76

Baldwin R, Taglioni D (2007) Trade effects of the euro: A comparison of estimators. J Econ Integration 780–818

Baldwin RE (2006) The euro's trade effects. ECB Working Paper No. 594. Available at SSRN: https://ssrn.com/abstract=886260 or https://doi.org/10.2139/ssrn.886260

Bergin PR, Lin C-Y (2012) The dynamic effects of a currency union on trade. J Int Econ 87(2):191–204

Bleaney M, Tian M (2012) Currency networks, bilateral exchange rate volatility and the role of the us dollar. Open Econ Rev 23:785–803

Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T (2006) Variable selection for propensity score models. Am J Epidemiol 163(12):1149–1156

Chintrakarn P (2008) Estimating the euro effects on trade with propensity score matching. Rev Int Econ 16(1):186–198
Conte M, Cotterlaz P, Mayer T (2021) The cepii gravity database. Paris, France, CEPII
Correia S, Guimarães P, Zylkin T (2019) Verifying the existence of maximum likelihood estimates for generalized linear models. arXiv preprint http://arxiv.org/abs/1903.01633
Correia S, Guimarães P, Zylkin T (2020) Fast poisson estimation with high-dimensional fixed effects. Stand Genomic Sci 20(1):95–115
Egger PH, Tarlea F (2015) Multi-way clustering estimation of standard errors in gravity models. Econ Lett 134:144–147
Felbermayr G, Groeschl J, Heiland I (2022) Complex europe: Quantifying the cost of disintegration. J Int Econ 138
Glick R, Rose AK (2002) Does a currency union affect trade? the time-series evidence. Eur Econ Rev 46(6):1125–1151
Glick R, Rose AK (2016) Currency unions and trade: A post-emu reassessment. Eur Econ Rev 87:78–91
Guimarães P Portugal P (2009) A simple feasible alternative procedure to estimate models with high-dimensional fixed effects
Head K, Mayer T (2014) Gravity equations: Workhorse, toolkit, and cookbook. In Handbook of international economics Volume4, pp. 131–195. Elsevier
Hirano K, Imbens GW (2001) Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Serv Outcomes Res Method 2(3):259–278
Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. Rev Econ Stat 86(1):4–29
Jordà Ò, Taylor AM (2016) The time for austerity: estimating the average treatment effect of fiscal policy. Econ J 126(590):219–255
Kelejian H, Tavlas GS, Petroulas P (2012) In the neighborhood: The trade effects of the euro in a spatial framework. Reg Sci Urban Econ 42(1–2):314–322
Kenen PB (2002) Currency unions and trade: Variations on themes by rose and persson. Reserve Bank of New Zealand Discussion Paper No. DP2002/08
Kopecky J (2023). Many unions, one estimate? disaggregating the currency union effect on trade. Emerging Markets Finance and Trade 1–23
Larch M, Wanner J, Yotov YV, Zylkin T (2019) Currency unions and trade: A ppml re-assessment with high-dimensional fixed effects. Oxford Bull Econ Stat 81(3):487–510
Lunceford JK, Davidian M (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med 23(19):2937–2960
Martínez-Zarzoso I (2019) The euro and the cfa franc: Evidence of sectoral trade effects. Open Econ Rev 30(3):483–504
Micco A, Stein E, Ordoñez G (2003) The currency union effect on trade: early evidence from emu. Economic policy 18(37):315–356
Millimet DL, Tchernis R (2009) On the specification of propensity scores, with applications to the analysis of trade policies. J Bus Econ Stat 27(3):397–415
Mundell RA (1961) A theory of optimum currency areas. Am Econ Rev 51(4):657–665
Nitsch V (2002) Honey, i shrunk the currency union effect on trade. The World Economy 25(4):457–457
O'Rourke KH Taylor AM (2013) Cross of euros. J Econ Perspect 167–191
Persson T (2001) Currency unions and trade: how large is the treatment effect? Economic policy 16(33):434–448
Polák P (2019) The euro's trade effect: a meta-analysis. J Econ Surv 33(1):101–124
Rose A (2002) Honey, the currency union effect on trade hasn't blown up. The World Economy 25(4):475–479
Rose AK (2000) One money, one market: the effect of common currencies on trade. Economic policy 15(30):08–45
Rose AK (2001) Currency unions and trade: the effect is large. Economic Policy 16(33):449–461
Rose AK (2017) Why do estimates of the emu effect on trade vary so much? Open Econ Rev 28(1):1–18
Silva JS, Tenreyro S (2006) The log of gravity. Rev Econ Stat 88(4):641–658
Silva JS, Tenreyro S (2011) Further simulation evidence on the performance of the poisson pseudo-maximum likelihood estimator. Econ Lett 112(2):220–222
Sloczyński T, Wooldridge JM (2018) A general double robustness result for estimating average treatment effects. Econometr Theor 34(1):112–133
Wooldridge JM (2007) Inverse probability weighted estimation for general missing data problems. Journal of econometrics 141(2):1281–1301